On-Device AI in Art Aura

How Apple Intelligence Powers a Smarter Art Inventory — Without Ever Leaving Your iPad

---

White Paper | February 2026

Art Aura by Republic

REPUBLIC

Executive Summary

Art Aura is an inventory management app for art collectors, galleries, and dealers. Starting with version 2.0, it uses Apple Intelligence — the AI system built into newer iPhones and iPads — to help users write professional artwork descriptions and search their collection using everyday language.

What makes this different from other AI-powered apps is simple: everything happens on the device itself. No artwork data is uploaded to the cloud. No subscription is required. No internet connection is needed. The AI runs directly on the chip inside your iPad or iPhone, using the same Neural Engine that powers Face ID and computational photography.

This paper explains, in plain language, how Art Aura uses Apple's on-device AI, why it matters for people who work with valuable art, and how the latest Apple hardware makes it all possible.

Table of Contents

1. The Problem: Writing About Art Is Hard

Anyone who manages an art collection knows the tedium of cataloging. Each artwork needs:

- A catalog description for records and insurance (2–3 polished sentences)
- An exhibition label for gallery walls (concise, 30–50 words)
- Historical context for provenance research and sales materials

Writing these well requires art-historical knowledge, consistent tone, and significant time. A gallery with 500 works might spend weeks just writing descriptions. And when you need to find “that large bronze sculpture we bought last year for under $10,000” — you’re stuck scrolling through lists or remembering exact titles.

Art Aura’s AI features solve both problems.

2. What Apple Intelligence Actually Is

Apple Intelligence is not a single technology — it’s a family of AI models that Apple has trained and optimized to run directly on their devices. Announced at WWDC 2025 as part of iOS 26, it gives app developers access to a language model (a program that understands and generates text) through a framework called Foundation Models.

Think of it this way:

Cloud AI (like ChatGPT)	Apple Intelligence
Your data goes to a server	Your data stays on your device
Requires internet	Works offline
Monthly subscription	Free with your device
General-purpose, massive model	Smaller, optimized for tasks
Runs on warehouse computers	Runs on the chip in your hand

Apple’s approach trades some of the raw power of cloud-based AI for something that matters deeply in the art world: absolute privacy and zero dependency on connectivity.

3. How Art Aura Uses On-Device AI

Art Aura integrates Apple Intelligence for two core features:

3.1 AI Description Generator

When viewing any artwork in your inventory, you can tap “Generate with AI.” The app sends the artwork’s basic metadata — title, artist, year, medium, and dimensions — to the on-device language model. Within seconds, it returns three professionally written texts:

Catalog Description

> *A professional 2–3 sentence description suitable for catalogs, insurance documentation, and collection records. Written in the measured, authoritative tone expected by the art world.*

Exhibition Label

> *A concise 30–50 word label in the style you'd see on a gallery wall next to the artwork. Informative but accessible.*

Historical Context

> *1–2 sentences placing the work within its art-historical moment — the movement, the artist's career phase, or cultural significance.*

You review each option and tap "Use This" to apply it to the artwork record. You remain in complete control; the AI suggests, you decide.

3.2 AI Tag Generation

Beyond descriptions, the AI can automatically suggest classification tags for each artwork:

- Style — The art movement (e.g., Impressionism, Abstract Expressionism)
- Subject — What the work depicts (e.g., landscape, portrait, still life)
- Color Palette — Dominant colors
- Mood — The emotional quality (e.g., contemplative, vibrant, somber)
- Keywords — 3–5 descriptive terms for search and organization

These tags make your collection instantly more searchable and better organized, without you having to manually classify each piece.

3.3 Smart Search

Instead of navigating filters and dropdowns, you simply type what you're looking for in plain language:

- "Available paintings under $10,000"
- "Large bronze sculptures"
- "Works by Anders Ström"
- "What's on consignment?"
- "Clients in New York"
- "Unpaid invoices from this year"

The AI parses your question into a structured query — understanding that "under $10,000" means a price filter, "available" means a status filter, and "paintings" should also match oil, acrylic, watercolor, gouache, tempera, and fresco — then returns the matching results instantly.

4. The Privacy Advantage

This is where on-device AI fundamentally changes the equation for anyone working with valuable art.

Why It Matters

Art collections represent significant financial and cultural value. Collection databases contain:

- Financial data — Purchase prices, sale amounts, insurance valuations
- Ownership records — Who owns what, provenance chains
- Location data — Where artworks are physically stored
- Client information — Collector names, contact details, buying patterns
- Business intelligence — Commission rates, gallery margins, market trends

Sending this information to a cloud AI service — even one with strong privacy policies — creates risk. Data breaches happen. Servers get subpoenaed. Terms of service change. Training data gets reused.

How Art Aura Eliminates That Risk

With Apple's on-device approach:

- No data leaves the device. When you ask Art Aura to generate a description for a $2 million painting, that request never touches the internet. The language model runs in a sandboxed process on the Neural Engine inside your iPad.
- No server to breach. There is no Art Aura AI server, no API endpoint, no database of your queries. The AI exists only as compiled code on the Apple silicon chip.
- No training on your data. Apple's Foundation Models framework explicitly prohibits using on-device interactions to train or improve the model. Your collection data is never fed back into any learning process.
- Works offline. At an art fair in Basel, a private viewing in a collector's home, or a storage facility with no WiFi — the AI works exactly the same as in your office.

This isn't a privacy policy you have to trust. It's a technical architecture that makes data leakage physically impossible.

5. How the Hardware Makes It Work

Running an AI language model on a handheld device — fast enough to feel instant — is an extraordinary engineering achievement. Here's how Apple's hardware makes it possible.

The Neural Engine

Every modern Apple chip contains a dedicated Neural Engine — specialized silicon designed exclusively for machine learning computations. Think of it as a separate brain within the chip that handles AI tasks, leaving the main processor free for everything else.

Chip	Neural Engine	TOPS*	Found In
A17 Pro	16 cores	35	iPhone 15 Pro / Pro Max
M1	16 cores	11	iPad Air 5th, iPad Pro 2021
M2	16 cores	15.8	iPad Air 6th, iPad Pro 2022
M4	16 cores	38	iPad Pro 2024
M5	16 cores	38+	iPad Pro 2025

**TOPS = Trillions of Operations Per Second*

To put 38 TOPS in perspective: generating a professional artwork description requires the model to evaluate billions of possible word combinations, score them for relevance and quality, and produce coherent text — all in under 30 seconds. The Neural Engine handles this workload while the main CPU and GPU remain free, so the app stays responsive the entire time.

Unified Memory Architecture

Apple silicon uses unified memory — the CPU, GPU, and Neural Engine all share the same pool of fast memory. This eliminates the bottleneck of copying data between different memory banks, which is how traditional computers work.

For Art Aura, this means:

- The language model loads directly into shared memory once
- The Neural Engine reads artwork metadata instantly, without waiting for data transfers
- Generated text flows back to the app with zero copy overhead
- The entire process feels like the AI is “thinking” for a moment, then responding

On a device with a discrete GPU and separate memory (like a traditional laptop), the same operation would require multiple data copies and could take significantly longer.

Memory Bandwidth

The speed at which data moves through the system matters as much as raw computation:

Chip	Memory Bandwidth
M1	68.25 GB/s
M2	100 GB/s
M4	120 GB/s
M5	120+ GB/s

Apple’s Foundation Models are quantized (compressed) to fit in limited device memory while maintaining quality. High memory bandwidth ensures the model’s parameters stream through the Neural Engine fast enough to produce text that feels near-instantaneous.

Thermal Design

Running AI workloads generates heat. Apple’s iPad Pro thermal design — with its large aluminum chassis acting as a heat sink — sustains Neural Engine performance for extended sessions. This matters when a gallery assistant is generating descriptions for dozens of artworks in a row, or when a dealer is running repeated searches during an art fair.

The iPhone 15 Pro, with its titanium frame and graphite thermal substrate, achieves similar sustained performance in a smaller form factor.

6. Smart Search: Ask Questions, Get Answers

Art Aura's Smart Search deserves special attention because it showcases a sophisticated use of on-device AI: turning human language into database queries.

How It Works

When you type "available paintings under $10,000" into Smart Search, here's what happens behind the scenes:

Step 1 — Intent Parsing

The on-device language model reads your query and produces a structured intent — a machine-readable breakdown of what you're looking for:

```
Entity: artwork
Filters:
  - status = available
  - medium contains "painting"
  - price < 10000
Sort: none specified (default order)
```

Step 2 — Medium Expansion

Art Aura knows that "painting" is a category, not just a medium. So it expands the search to include: oil, acrylic, watercolor, watercolour, gouache, tempera, fresco, and encaustic. A search for "sculpture" similarly expands to bronze, marble, stone, wood, ceramic, glass, metal, clay, plaster, and resin.

Step 3 — Filtering

The structured filters are applied against your collection. Price comparisons handle currency symbols and formatting. Status codes are matched. String comparisons support international characters (important for artists with names like Anders Ström or François Müller).

Step 4 — Results

Matching artworks appear instantly, with the option to view debug information showing exactly which filters were applied.

Why On-Device Matters for Search

Cloud-based search would require uploading your entire inventory to a server for querying. On-device search means:

- Your collection data stays local
- Results appear in milliseconds (no network round-trip)
- It works without internet
- Search patterns can't be profiled or sold

7. Graceful Degradation: It Works for Everyone

Not every device supports Apple Intelligence. Art Aura handles this gracefully with a layered approach:

For AI-Capable Devices (M1+ iPad, iPhone 15 Pro+)

Full Apple Intelligence integration. Natural language search is parsed by the on-device model. Descriptions are generated by AI. Tags are suggested automatically.

For Other Devices

Art Aura falls back to a keyword-based parser that still handles:

- Price ranges (“under $10k”, “over €5,000”)
- Status keywords (“available”, “sold”, “reserved”)
- Medium matching with the same expansion rules
- Size keywords (“large”, “small”)
- Year and decade parsing (“1960s”, “after 2010”)
- Artist name detection with Unicode support
- Client city matching
- Invoice status filters
- Sorting (“most expensive”, “recent”, “oldest”)

The fallback parser covers the vast majority of real-world queries. The AI layer adds nuance — understanding ambiguous phrasing, implied filters, and complex multi-part questions — but the app remains fully functional without it.

Settings Control

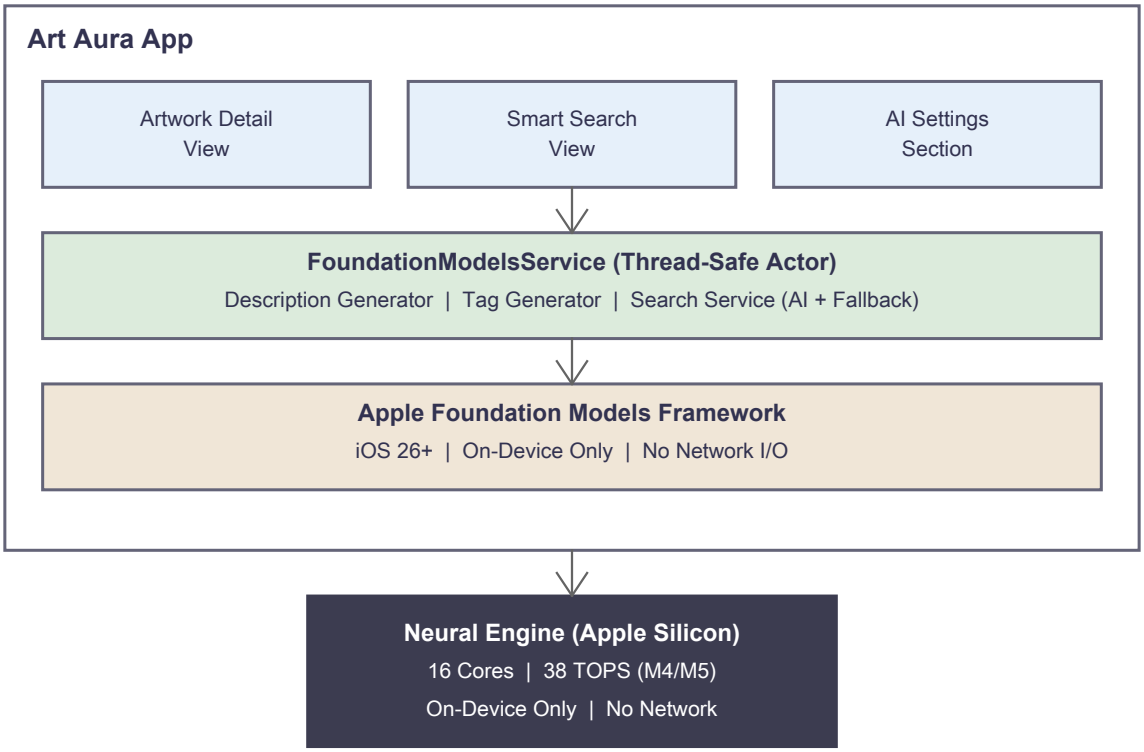
Users can check AI availability in the app’s settings, where a clear status indicator shows:

- Available (green checkmark) — AI features are ready
- Unavailable (gray X) — With a human-readable explanation of why

An “Enable AI Features” toggle lets users opt out even on supported devices, respecting user autonomy.

8. Technical Architecture (Simplified)

For those interested in how the pieces fit together, here's a simplified view of the architecture:



Key Design Decisions

Actor-Based Concurrency

The AI service uses Swift's actor model — a modern concurrency primitive that guarantees thread safety. Multiple parts of the app can request AI generation simultaneously without conflicts or crashes. This is like having a single, well-organized assistant who handles one request at a time, queuing the rest.

Structured Output

Rather than generating free-form text and hoping it's usable, Art Aura uses Apple's @Generable protocol. This tells the AI model exactly what shape the output should take — "give me a catalog description, an exhibition label, and historical context as three separate fields." The result is always structured and ready to use, never garbled or off-format.

Conditional Compilation

The AI code is wrapped in compile-time checks and runtime availability checks. This means the same app binary runs on every supported device — AI features simply activate when the hardware and software support them.

Session Reuse

The language model session is created once and reused across multiple requests. This avoids the overhead of reloading the model for each query, making the second and subsequent AI requests significantly faster than the first.

9. Frequently Asked Questions

Does Art Aura send my collection data to any server?

No. All AI processing runs on the Neural Engine inside your device. Art Aura has no AI server, no cloud processing, and no telemetry. Your data physically cannot leave the device through the AI feature.

Do I need an internet connection for AI features?

No. Once Apple Intelligence is set up on your device (a one-time download of 2–3 GB over WiFi), all AI features work completely offline.

Does it cost extra?

No. On-device AI is included with your Apple device and iOS 26. There are no API fees, no token limits, and no subscriptions for the AI features in Art Aura.

Which devices support it?

iPad Pro with M1, M2, M4, or M5 chip. iPad Air with M1 or M2 chip. iPhone 15 Pro and iPhone 15 Pro Max. Devices without these chips still get full app functionality — just without the AI description generator and AI-powered search parsing.

How accurate are the generated descriptions?

The AI produces professional-quality text based on the metadata you provide. The more complete your artwork record (title, artist, year, medium, dimensions), the better the output. You always review and approve before anything is saved.

Can I turn off AI features?

Yes. A toggle in Settings lets you disable AI features entirely, even on supported devices.

How long does generation take?

Typically under 30 seconds on M-series devices. The first request after launching the app may take slightly longer as the model session initializes. Subsequent requests are faster.

10. Conclusion

Art Aura's integration of Apple Intelligence represents a practical, privacy-first approach to AI in specialized professional software. By running entirely on-device:

- Collectors get professional descriptions without hiring a cataloger for every piece
- Galleries can search their inventory naturally, in the language they already use
- Dealers can work at art fairs and private viewings without depending on connectivity
- Everyone benefits from AI without surrendering sensitive collection data to cloud services

The latest Apple hardware — with its dedicated Neural Engine, unified memory, and high bandwidth — transforms what would have been a cloud-dependent feature into something that runs in the palm of your hand. No accounts to create. No terms of service to accept. No data to worry about.

The AI is in the device. Your data stays with you.